

A Tight Upper Bound for the Third-Order Asymptotics of Discrete Memoryless Channels

Marco Tomamichel^{1,*} and Vincent Y. F. Tan^{2,3,†}

¹*Centre for Quantum Technologies, National University of Singapore*

²*Institute for Infocomm Research (I²R), A*STAR, Singapore*

³*Department of Electrical and Computer Engineering, National University of Singapore*

This paper shows that the ε -error capacity (average error probability) for n uses of a discrete memoryless channel is upper bounded by the normal approximation plus a term that does not exceed $\frac{1}{2} \log n + O(1)$ if the ε -dispersion of the channel is positive. If the ε -dispersion vanishes, the ε -error capacity is upper bounded by the asymptotic capacity plus a constant term, unless the channel is exotic and $\varepsilon \geq \frac{1}{2}$.

I. INTRODUCTION

The primary information-theoretic task in point-to-point channel coding is the characterization of the maximum rate of communication over n independent uses of a noisy channel W . We are concerned in this paper with *discrete memoryless channels* (DMCs). Let $M^*(W^n, \varepsilon)$ resp. $M_{\max}^*(W^n, \varepsilon)$ denote the maximum size of a length- n block code for DMC W having *average* resp. *maximal* error probability no larger than $\varepsilon \in (0, 1)$. Shannon's *noisy-channel coding theorem* [1] and Wolfowitz's *strong converse* [2] state that for every $\varepsilon \in (0, 1)$,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log M^*(W^n, \varepsilon) = C \quad \text{bits/channel use,}$$

where $C := \max_P I(P, W)$ is the *channel capacity*. Since the mid-1960s, there has been interest in determining finer asymptotic characterizations of the coding theorem. This is useful because such an analysis provides key insights into the amount of backoff from channel capacity for block codes of finite length n . In particular, Strassen in 1964 [3] showed using normal approximations that the asymptotic expansion of $\log M_{\max}^*(W^n, \varepsilon)$ satisfies

$$\log M_{\max}^*(W^n, \varepsilon) = nC + \sqrt{nV_\varepsilon} \Phi^{-1}(\varepsilon) + \rho_n, \quad (1)$$

where $\rho_n = O(\log n)$, V_ε is the ε -*channel dispersion* [4, 5] and $\Phi(\cdot)$ is the Gaussian cumulative distribution function. These quantities will be defined precisely in Section II A. In fact, this asymptotic expansion also holds for $M^*(W^n, \varepsilon)$ [4, Eqs. (284)-(286)] and implies that if an error probability of ε is tolerable, the backoff from channel capacity C at finite blocklength n is roughly $\sqrt{V_\varepsilon/n} \Phi^{-1}(\varepsilon)$. There have been several recent refinements to and extensions of Strassen's normal approximation in (1), most prominently by Hayashi [6] and Polyanskiy-Poor-Verdú (PPV) [4]. Strassen's normal approximation has also been shown to hold for many other classes of channels such as the additive white Gaussian noise channel [4-6].

Despite these impressive advances in the fundamental limits of channel coding, the third-order term ρ_n is not well understood. Indeed, Hayashi in the conclusion of his paper [6] mentions that

“... the third-order coding rate is expected but appears difficult. The second order is the order \sqrt{n} , and it is not clear whether the third order is a constant order or the order $\log n$ ”

* cqtmarco@nus.edu.sg

† vtan@nus.edu.sg

What we do know is that for the binary symmetric channel (BSC), $\rho_n = \frac{1}{2} \log n + O(1)$ [4, Thm. 52] and for the binary erasure channel (BEC), $\rho_n = O(1)$ [4, Thm. 53]. More generally, there are classes of channels for which we have bounds on ρ_n [5, Sec. 3.4.5]. For lower bounds (achievability), if we consider DMCs W with positive capacity and all elements of the stochastic matrix W are positive, $\rho_n \geq \frac{1}{2} \log n + O(1)$ [5, Cor. 54]. For upper bounds (converse), if we restrict our attention to so-called *weakly input-symmetric* DMCs [5, Def. 9], $\rho_n \leq \frac{1}{2} \log n + O(1)$ [5, Thm. 55]. For *constant-composition codes* [7], it was shown [8] using strong large-deviation techniques [9, Sec. XVI.7] that, under some regularity assumptions, $\rho_n = \frac{1}{2} \log n + O(1)$. Recall that a constant-composition code is one where all the codewords are of the same *empirical distribution* or *type*. It is also claimed that the same holds for a more general class of DMCs in [10]. Our results generalize the converses in [8, 10].

This paper strengthens the upper bound (converse) on the third-order term ρ_n . For all DMCs whose ε -dispersions are positive, we show that

$$\log M^*(W^n, \varepsilon) \leq nC + \sqrt{nV_\varepsilon} \Phi^{-1}(\varepsilon) + \frac{1}{2} \log n + O(1), \quad (2)$$

If the ε -dispersion vanishes, the corresponding bound is $\log M^*(W^n, \varepsilon) \leq nC + O(1)$, unless the DMC is exotic [4, Thm. 48] and $\varepsilon \geq \frac{1}{2}$. If the DMC is exotic and $\varepsilon = \frac{1}{2}$, we show that $\log M^*(W^n, \frac{1}{2}) \leq nC + \frac{1}{2} \log n + O(1)$. If the DMC is exotic and $\varepsilon > \frac{1}{2}$, $\log M^*(W^n, \varepsilon) \leq nC + O(n^{\frac{1}{3}})$, a result by PPV [4, Thm. 48]. Hence, for the rather general class of DMCs with positive ε -dispersion, the third-order term is $\rho_n \leq \frac{1}{2} \log n + O(1)$. We may thus dispense with the assumption that W is weakly input-symmetric [5, Def. 9].

The typical way [3–7] to upper bound $M^*(W^n, \varepsilon)$ is to first do the same for the maximum size of a constant-composition code under the maximum error probability formulation, i.e., $M_{\max}^*(W^n, \varepsilon)$. Such a bound can be proved using either the meta-converse [4, Thm. 31] or tight bounds on the type-II error probability in a simple binary hypothesis test [3, Thm. 1.1]. By the type-counting lemma [7, Lem. 2.2], every length- n block code can be partitioned into no more than $(n+1)^{|\mathcal{X}|-1}$ constant-composition subcodes. This leads to the rather conservative bound [3, Eq. (4.29)] [4, Eq. (279)]

$$\log M_{\max}^*(W^n, \varepsilon) \leq nC + \sqrt{nV_\varepsilon} \Phi^{-1}(\varepsilon) + \left(|\mathcal{X}| - \frac{1}{2}\right) \log n + O(1). \quad (3)$$

Subsequently, by expurgating bad codewords (see [4, Eqs. (284)–(286)]), we can conclude that the same upper bound holds for $M^*(W^n, \varepsilon)$. We adopt a different approach for the proof of our main result in (2) and work with $M^*(W^n, \varepsilon)$ directly. In a nutshell, we generalize the converse technique in Wang-Colbeck-Renner [11] and Wang-Renner [12], exploit the link [13, Lem. 12] between the ε -hypothesis testing relative entropy [14] and the relative entropy information spectrum [15, Ch. 4] and carefully weigh the contributions of each input type for a general (non-constant-composition) code by constructing an appropriate ϵ -net for the output probability simplex. The last step, which replaces the use of the type-counting lemma, allows us to bound the effect of different input types with the $O(1)$ term in (2).

Note that unlike in (3), the third-order term in our upper bound in (2) is independent of $|\mathcal{X}|$. This is intuitive upon doing the following thought experiment. Let n be a large even integer and consider using transmitting information across n uses of a DMC $W : \mathcal{X} \rightarrow \mathcal{Y}$. Clearly, the same amount of information can be transmitted through $\frac{n}{2}$ uses of the product channel $W^2 : \mathcal{X}^{\times 2} \rightarrow \mathcal{Y}^{\times 2}$, where $W^2(y, y'|x, x') := W(y|x)W(y'|x')$. The capacity and the dispersion of W^2 are respectively twice the capacity and the dispersion of W so the normal approximation terms for n uses of W and $\frac{n}{2}$ uses of W^2 are identical. If the coefficient of the third-order logarithmic term *were* dependent on the size of the input alphabet, say via some

function $g(|\mathcal{X}|)$, then for the first experiment, $\rho_n = g(|\mathcal{X}|) \log n + O(1)$ while for the second experiment, $\rho_n = g(|\mathcal{X}|^2) \log(\frac{n}{2}) + O(1) = g(|\mathcal{X}|^2) \log n + O(1)$. Thus, at least at an intuitive level, we expect that $g(|\mathcal{X}|)$ is independent of $|\mathcal{X}|$.

II. NOTATION AND PRELIMINARIES

A. Discrete Memoryless Channels

As mentioned in the Introduction, we consider *discrete memoryless channels* (DMCs), which are characterized by two finite sets, the input alphabet \mathcal{X} and the output alphabet \mathcal{Y} , and a stochastic matrix W , where $W(y|x)$ denotes the probability that the output $y \in \mathcal{Y}$ occurs given input $x \in \mathcal{X}$. The set of probability distributions on \mathcal{X} is denoted $\mathcal{P}(\mathcal{X})$. For any probability distribution $P \in \mathcal{P}(\mathcal{X})$, we denote by $P \times W : (x, y) \mapsto P(x)W(y|x)$ the joint distribution of inputs and outputs of the channel, and by $PW : y \mapsto \sum_x P(x)W(y|x)$ its marginal on \mathcal{Y} . Finally, $W(\cdot|x)$ denotes the distribution on \mathcal{Y} if the input is fixed to x .

Given two probability distributions $P, Q \in \mathcal{P}(\mathcal{X})$, we call the random variable $\log \frac{P(X)}{Q(X)}$ where X has distribution P the *log-likelihood ratio* of P and Q . Its mean is the *relative entropy*

$$D(P\|Q) := \mathbb{E}_P \left[\log \frac{P}{Q} \right] = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}$$

and $D(W\|Q|P) := \sum_x P(x)D(W(\cdot|x)\|Q)$ is the *conditional information divergence*. The *mutual information* is $I(P, W) := D(W\|PW|P)$. Moreover,

$$C(W) := \max_{P \in \mathcal{P}(\mathcal{X})} I(P, W) \quad \text{and} \quad \Pi(W) := \{P \in \mathcal{P}(\mathcal{X}) \mid I(P, W) = C(W)\}$$

are the *capacity* and the set of *capacity-achieving input distributions* (CAIDs), respectively.¹ The set of CAIDs is convex and compact in $\mathcal{P}(\mathcal{X})$. The unique [16, Cor. 2 to Thm. 4.5.1] *capacity-achieving output distribution* (CAOD) is denoted as Q^* and $Q^* = PW$ for all $P \in \Pi$. Furthermore, it satisfies $Q^*(y) > 0$ for all $y \in \mathcal{Y}$ [16, Cor. 1 to Thm. 4.5.1], where we assume that all outputs are accessible.

The variance of the log-likelihood ratio of P and Q is the *divergence variance*

$$V(P\|Q) := \mathbb{E}_P \left[\left(\log \frac{P}{Q} - D(P\|Q) \right)^2 \right].$$

We also define the *conditional divergence variance* $V(W\|Q|P) := \sum_x P(x)V(W(\cdot|x)\|Q)$ and the *conditional information variance* $V(P, W) := V(W\|PW|P)$. Note that $V(P, W) = V(P \times W\|P \times PW)$ for all $P \in \Pi$ [4, Lem. 62]. The ε -*channel dispersion*² [4, Def. 2] is an operational quantity that was shown [4, Eq. (223)] to be equal to

$$V_\varepsilon(W) := \begin{cases} V_{\min} & \text{if } \varepsilon < \frac{1}{2} \\ V_{\max} & \text{if } \varepsilon \geq \frac{1}{2} \end{cases}, \quad \text{where} \quad V_{\min} := \min_{P \in \Pi} V(P, W) \quad \text{and} \quad V_{\max} := \max_{P \in \Pi} V(P, W).$$

Furthermore, a channel is called *exotic* [4, before Thm. 48] if $V_{\max} = 0$ and there exists a symbol $x_0 \in \mathcal{X}$ such that $D(W(\cdot|x_0)\|Q^*) = C$ and $V(W(\cdot|x_0)\|Q^*) > 0$.³

¹ We often drop the dependence on W if it is clear from context.

² Notice that for $\varepsilon = \frac{1}{2}$, we set $V_\varepsilon = V_{\max}$. This is somewhat unconventional; cf. [4, Thm. 48]. However, doing so ensures that Theorem 1 can be stated compactly. Nonetheless, from the viewpoint of the normal approximation, it is immaterial how we choose $V_{\frac{1}{2}}$ since $\Phi^{-1}(\frac{1}{2}) = 0$ (cf. [4, after Eq. (280)]).

³ Note that this symbol must satisfy $P(x_0) = 0$ for any $P \in \Pi$, as otherwise V_{\max} would not vanish.

For later reference, we also define the *third absolute moment of the log-likelihood ratio*,

$$T(P\|Q) := \mathbb{E}_P \left[\left| \log \frac{P}{Q} - D(P\|Q) \right|^3 \right]$$

and $T(W\|Q|P) := \sum_x P(x)T(W(\cdot|x)|Q)$.

We employ the cumulative distribution function of the standard normal distribution

$$\Phi(a) := \int_{-\infty}^a \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) dx$$

and define its inverse as $\Phi^{-1}(\varepsilon) := \sup\{a \in \mathbb{R} \mid \Phi(a) \leq \varepsilon\}$, which evaluates to the usual inverse for $0 < \varepsilon < 1$ and continuously extended to take values $\pm\infty$ outside that range.

For a sequence $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^{\times n}$, we denote by $P_{\mathbf{x}} \in \mathcal{P}(\mathcal{X})$ the probability distribution given by the relative frequencies of \mathbf{x} , i.e. $P_{\mathbf{x}}(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{x_i=x\}}$. This probability distribution $P_{\mathbf{x}}$ is also known as the *empirical distribution* or the *type* [7] of \mathbf{x} . The set of all such distributions is denoted as $\mathcal{P}_n(\mathcal{X}) = \bigcup_{\mathbf{x}} \{P_{\mathbf{x}}\}$ and satisfies $|\mathcal{P}_n(\mathcal{X})| \leq (n+1)^{|\mathcal{X}|-1}$.

B. Codes and ε -Error Capacity

A *code* \mathcal{C} for a channel is defined by the triple $\{\mathcal{M}, e, d\}$, where \mathcal{M} is a set of messages, $e : \mathcal{M} \rightarrow \mathcal{X}$ an encoding function and $d : \mathcal{Y} \rightarrow \mathcal{M}$ a decoding function. We write $|\mathcal{C}| = |\mathcal{M}|$ for the cardinality of the message set. We define the *average error probability* of a code \mathcal{C} for the channel W as

$$p_{\text{err}}(\mathcal{C}, W) := \Pr[M \neq M'] = 1 - \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} W(d^{-1}(m)|e(m))$$

where the distribution over messages P_M is assumed to be uniform on \mathcal{M} ,

$$M \xrightarrow{e} X \xrightarrow{W} Y \xrightarrow{d} M'$$

forms a Markov chain, and M' thus denotes output of the decoder. The *one-shot ε -error capacity* of the channel W is then defined as

$$M^*(W, \varepsilon) := \max \{m \in \mathbb{N} \mid \exists \mathcal{C} : |\mathcal{C}| = m \wedge p_{\text{err}}(\mathcal{C}, W) \leq \varepsilon\}.$$

We are also interested in the ε -error capacity for $n \geq 1$ uses of a memoryless channel. For this purpose, we consider the channel W^n , defined by the stochastic matrix $W^n(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^n W(y_i|x_i)$, where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$ are strings of length n of symbols $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, respectively. Then, the *blocklength n , ε -error capacity* of the channel W is denoted as $M^*(W^n, \varepsilon)$.

III. MAIN RESULT AND PROOF

Let us reiterate our main result. The various cases are illustrated diagrammatically in Fig. 1.

Theorem 1. *For every DMC W and ε with $V_\varepsilon > 0$, the blocklength n , ε -error capacity satisfies*

$$\log M^*(W^n, \varepsilon) \leq nC + \sqrt{nV_\varepsilon} \Phi^{-1}(\varepsilon) + \frac{1}{2} \log n + O(1).$$

If $V_\varepsilon = 0$, we have $\log M^(W^n, \varepsilon) \leq nC + O(1)$, unless the channel is exotic and $\varepsilon \geq \frac{1}{2}$.*

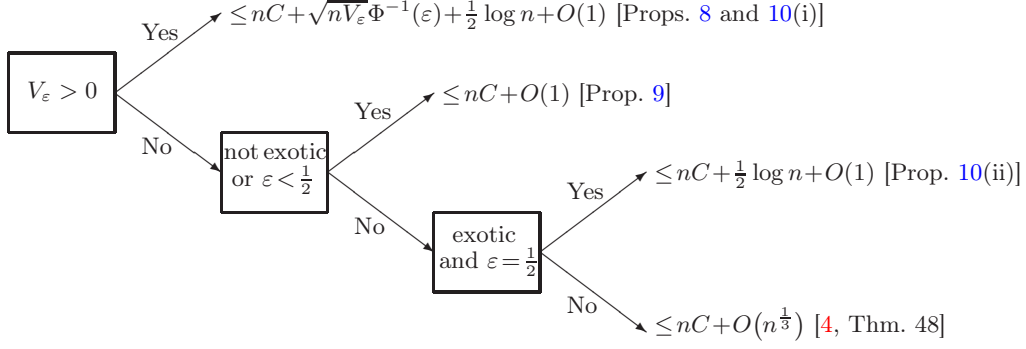


FIG. 1. Illustration of the various cases of Theorem 1 and the proof structure in Section III E

Remark 1. For exotic channels and $\varepsilon > \frac{1}{2}$, PPV showed the converse bound $\log M_{\max}^*(W^n, \varepsilon) \leq nC + O(n^{\frac{1}{3}})$ [4, Thm. 48]. It was also shown, via an example, that the $O(n^{\frac{1}{3}})$ term cannot be improved in general [4, App. H].

Remark 2. The $\varepsilon = \frac{1}{2}$ case needs to be treated with care. For exotic channels and $\varepsilon = \frac{1}{2}$ (so $V_\varepsilon = 0$), we show that $\log M^*(W^n, \varepsilon) \leq nC + \frac{1}{2} \log n + O(1)$. In fact, for all DMCs W with $V_{\min} = 0$ and $\varepsilon = \frac{1}{2}$, we show that $\log M^*(W^n, \varepsilon) \leq nC + \frac{1}{2} \log n + O(1)$. See Proposition 10. If $V_{\max} > 0$, the latter statement concurs with the positive ε -dispersion case of Theorem 1.

Remark 3. From the preceding statements, we see that for DMCs with $V_{\min} = 0$ and $V_{\max} > 0$, the third-order term “jumps” from 0 to $\frac{1}{2} \log n$ when $\varepsilon \uparrow \frac{1}{2}$. This is because we do not investigate the dependence of the constant term on ε . If we did, for the case $V_{\min} = 0$, $V_{\max} > 0$ and $\varepsilon = (\frac{1}{2})^-$, we would notice that the constant term diverges as $\varepsilon \uparrow \frac{1}{2}$.

In light of the existing results on ρ_n (in the Introduction and [5, Sec. 3.4.5]), the third order term is the best possible unless we impose further assumptions on W .

The proof consists of five parts, each detailed in one of the following subsections. In the first subsection, we introduce two entropic quantities, the hypothesis testing divergence [11–14] and a quantity related to the information (or divergence) spectrum [15, Ch. 4]. We state and prove some useful properties we need later. In the second subsection, we derive a converse bound, valid for general DMCs, that involves a minimization over output distributions and maximization over input symbols. In the third subsection, we choose an appropriate output distribution for use in the general converse bound. In the fourth subsection, we state and prove some continuity properties of information measures around the CAIDs and the unique CAOD. Finally, the fifth subsection contains the proof of our main result.

A. Hypothesis Testing and the Information Spectrum

We use the following divergence [11–14], which is closely related to binary hypothesis testing. Let $\varepsilon \in (0, 1)$ and let $P, Q \in \mathcal{P}(\mathcal{Z})$, where \mathcal{Z} is finite. We consider binary (probabilistic) hypothesis tests $\xi : \mathcal{Z} \rightarrow [0, 1]$ and define the ε -hypothesis testing divergence

$$D_h^\varepsilon(P\|Q) := \sup \left\{ R \in \mathbb{R} \mid \exists \xi : \mathbb{E}_Q[\xi(Z)] \leq (1 - \varepsilon) \exp(-R) \wedge \mathbb{E}_P[\xi(Z)] \geq 1 - \varepsilon \right\}.$$

Note that $D_h^\varepsilon(P\|Q) = -\log \frac{\beta_{1-\varepsilon}(P, Q)}{1-\varepsilon}$ where β_α is defined in PPV [4, Eq. (100)]. It is easy to see that $D_h^\varepsilon(P\|Q) \geq 0$, where the lower bound is achieved if and only if $P = Q$ and $D_h^\varepsilon(P\|Q)$ diverges if P and Q are orthogonal. It satisfies a data-processing inequality [11]

$$D_h^\varepsilon(P\|Q) \geq D_h^\varepsilon(PW\|QW) \quad \text{for all channels } W \text{ from } \mathcal{Z} \text{ to } \mathcal{Z}'.$$

When evaluated for independent and identical distributions (i.i.d.), its asymptotic expansion in the first order is determined by the Chernoff-Stein Lemma [7, Cor. 1.2], yielding $D_h^\varepsilon(P^{\times n} \| Q^{\times n}) = nD(P \| Q) + o(n)$ for any $\varepsilon \in (0, 1)$. This analysis was tightened by Strassen [3, Thm. 3.1] and he showed that

$$D_h^\varepsilon(P^{\times n} \| Q^{\times n}) = nD(P \| Q) + \sqrt{nV(P \| Q)}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + O(1).$$

The following quantity, which characterizes the distribution of the log-likelihood ratio and is known as the *relative entropy information spectrum* or the *divergence spectrum* [15, Ch. 4], is sometimes easier to manipulate and evaluate.

$$D_s^\varepsilon(P \| Q) := \sup \left\{ R \in \mathbb{R} \mid \Pr_P \left[\log \frac{P}{Q} \leq R \right] \leq \varepsilon \right\}.$$

It is intimately related to the ε -hypothesis testing divergence, as the following lemma shows.

Lemma 2. *For any $\delta \in (0, 1 - \varepsilon)$, we have*

$$D_s^\varepsilon(P \| Q) - \log \frac{1}{1 - \varepsilon} \leq D_h^\varepsilon(P \| Q) \leq D_s^{\varepsilon + \delta}(P \| Q) + \log \frac{1 - \varepsilon}{\delta}. \quad (4)$$

These relations follow from standard arguments relating binary hypothesis testing and the log-likelihood test to the relative entropy information spectrum. In [13, Lem. 12], an analogue of the above lemma is shown for the strictly more general non-commutative case. For completeness we show the second inequality, which we will employ later.

Proof of Second Inequality in (4). If $D_h^\varepsilon(P \| Q)$ is infinite, P is not absolutely continuous with respect to Q and it is easy to see that $D_s^{\varepsilon + \delta}(P \| Q)$ is also infinite. Hence, the second inequality in (4) trivially holds. We thus consider the case where $D_h^\varepsilon(P \| Q)$ is finite, and fix any optimal test ξ for $D_h^\varepsilon(P \| Q)$. Set $R^* := D_h^\varepsilon(P \| Q) + \log \frac{\delta}{1 - \varepsilon}$. We find

$$\begin{aligned} \Pr_P \left[\log \frac{P}{Q} > R^* \right] &= \sum_{z \in \mathcal{Z}} P(z) 1_{\{P(z) > \exp(R^*)Q(z)\}} \\ &\geq \sum_{z \in \mathcal{Z}} (P(z) - \exp(R^*)Q(z)) 1_{\{P(z) > \exp(R^*)Q(z)\}} \\ &\geq \sum_{z \in \mathcal{Z}} (P(z) - \exp(R^*)Q(z)) \xi(z) \\ &= \mathbb{E}_P[\xi(Z)] - \exp(R^*) \mathbb{E}_Q[\xi(Z)] \\ &\geq 1 - \varepsilon - \delta. \end{aligned}$$

In the last step we used the fact that ξ is an optimal test, which implies that $\mathbb{E}_P[\xi(Z)] \geq 1 - \varepsilon$ and $\mathbb{E}_Q[\xi(Z)] \leq (1 - \varepsilon) \exp(-D_h^\varepsilon(P \| Q))$. Thus, $D_s^{\varepsilon + \delta}(P \| Q) \geq R^*$, concluding the proof. \square

We can give an upper bound on $D_s^\varepsilon(P \| Q)$ if Q is a convex combination of distributions.

Lemma 3. *Let $P \in \mathcal{P}(\mathcal{Z})$ and $Q = \sum_{i \in \mathcal{I}} q(i)Q^i$ with $Q^i \in \mathcal{P}(\mathcal{Z})$ and $q \in \mathcal{P}(\mathcal{I})$ and \mathcal{I} is some countable index set. Then,*

$$D_s^\varepsilon(P \| Q) \leq \inf \{ D_s^\varepsilon(P \| Q^i) - \log q(i) \}_{i \in \mathcal{I}}$$

Proof. Note that for all $z \in \mathcal{Z}$, for all $i \in \mathcal{I}$, we have

$$\log \frac{P(z)}{Q(z)} = \log \frac{P(z)}{\sum_j q(j)Q^j(z)} \leq \log \frac{P(z)}{q(i)Q^i(z)} = \log \frac{P(z)}{Q^i(z)} - \log q(i).$$

Hence,

$$\Pr_P \left[\log \frac{P}{Q} \leq R \right] \geq \Pr_P \left[\log \frac{P}{Q^i} \leq R + \log q(i) \right]$$

and, relaxing the optimization in the definition of D_s^ε , we get $D_s^\varepsilon(P\|Q) \leq D_s^\varepsilon(P\|Q^i) - \log q(i)$ as desired. \square

The following property will be particularly useful, as it allows us to bound the log-likelihood ratio of the input-output behavior of two channels in terms of the log-likelihood ratio evaluated for a single input symbol.

Lemma 4. *Let $P \in \mathcal{P}(\mathcal{X})$ and let V, W be channels from \mathcal{X} to \mathcal{Y} . Then,*

$$D_s^\varepsilon(P \times W \| P \times V) \leq \sup_{x: P(x) > 0} D_s^\varepsilon(W(\cdot|x) \| V(\cdot|x)).$$

Proof. We first note that the log-likelihood ratio takes on the form

$$\log \frac{P \times W}{P \times V} : (x, y) \mapsto \log \frac{P(x)W(y|x)}{P(x)V(y|x)} = \log \frac{W(y|x)}{V(y|x)},$$

for every $(x, y) \in \mathcal{X} \times \mathcal{Y}$ satisfying $P(x) > 0$. Now, we may write

$$\begin{aligned} R^* = D_s^\varepsilon(P \times W \| P \times V) &= \sup \left\{ R \in \mathbb{R} \mid \Pr_{P \times W} \left[\log \frac{P \times W}{P \times V} \leq R \right] \leq \varepsilon \right\} \\ &= \sup \left\{ R \in \mathbb{R} \mid \sum_{x: P(x) > 0} P(x) \Pr_{W(\cdot|x)} \left[\log \frac{W(\cdot|x)}{V(\cdot|x)} \leq R \right] \leq \varepsilon \right\}. \end{aligned}$$

Inspecting this expression, for any $\varphi > 0$, we find at least one $x^* \in \mathcal{X}$ such that

$$P(x^*) > 0 \quad \text{and} \quad \Pr_{W(\cdot|x^*)} \left[\log \frac{W(\cdot|x^*)}{V(\cdot|x^*)} \leq R^* - \varphi \right] \leq \varepsilon.$$

Hence, $D_s^\varepsilon(W(\cdot|x^*) \| V(\cdot|x^*)) \geq R^* - \varphi$, which implies the lemma as φ is arbitrary. \square

The distribution of the log-likelihood ratio has the following asymptotic expansions for not necessarily identical product distributions.

Lemma 5. *Let $P_i, Q \in \mathcal{P}(\mathcal{Z})$ be such that $P_i \ll Q$ for all i in some finite set \mathcal{I} . We consider a sequence of distributions P_{i_k} indexed by (i_1, i_2, \dots, i_n) where $i_k \in \mathcal{I}$ for each $1 \leq k \leq n$. Define*

$$D_n := \frac{1}{n} \sum_{k=1}^n D(P_{i_k} \| Q), \quad V_n := \frac{1}{n} \sum_{k=1}^n V(P_{i_k} \| Q), \quad \text{and} \quad T_n := \frac{1}{n} \sum_{k=1}^n T(P_{i_k} \| Q).$$

If $V_n > 0$, then we have the Berry-Esséen-type bound

$$D_s^\varepsilon(P_{i_1} \times \dots \times P_{i_n} \| Q^{\times n}) \leq nD_n + \sqrt{nV_n} \Phi^{-1} \left(\varepsilon + \frac{6T_n}{\sqrt{nV_n^3}} \right).$$

In any case, we have the Chebyshev-type bound

$$D_s^\varepsilon(P_{i_1} \times \dots \times P_{i_n} \| Q^{\times n}) \leq nD_n + \sqrt{\frac{nV_n}{1-\varepsilon}}. \quad (5)$$

Proof. We consider the cumulative distribution of the random variable $S_n := \sum_k \log P_{i_k}(X_{i_k}) - \log Q(X_{i_k})$ where each X_{i_k} has distribution P_{i_k} . The random variable S_n has mean nD_n and variance nV_n . The general case, Eq. (5), is shown using Chebyshev's inequality, which yields

$$\varepsilon \geq \Pr \left[\sum_k \log \frac{P_{i_k}}{Q} \leq R \right] \geq 1 - \frac{nV_n}{(R - nD_n)^2} \quad \text{for } R > nD_n$$

Hence, restricting to $R > nD_n$ and relaxing the condition of the supremum, we find

$$D_s^\varepsilon(P_{i_1} \times \dots \times P_{i_n} \| Q^{\times n}) \leq \sup \left\{ R > nD_n \mid 1 - \frac{nV_n}{(R - nD_n)^2} \leq \varepsilon \right\} = nD_n + \sqrt{\frac{nV_n}{1 - \varepsilon}}.$$

Furthermore, if $V_n \geq V_- > 0$, the Berry-Esséen theorem [9, Sec. XVI.5] states that

$$\left| \Pr \left[\sum_k \log \frac{P_{i_k}}{Q} \leq R \right] - \Phi \left(\frac{R - nD_n}{\sqrt{nV_n}} \right) \right| \leq \frac{6T_n}{\sqrt{nV_n^3}}.$$

Hence, we obtain

$$D_s^\varepsilon(P_{i_1} \times \dots \times P_{i_n} \| Q^{\times n}) \leq nD_n + \sqrt{nV_n} \Phi^{-1} \left(\varepsilon + \frac{6T_n}{\sqrt{nV_n^3}} \right),$$

which concludes the proof. \square

B. Converse Bounds on General Channels

Here, we give a new converse bound on the code size for general channels.

Proposition 6. *Let $\varepsilon \in (0, 1)$ and let W be any channel from \mathcal{X} to \mathcal{Y} . Then, for any $\delta \in (0, 1 - \varepsilon)$, we have*

$$\log M^*(W, \varepsilon) \leq \inf_{Q \in \mathcal{P}(\mathcal{Y})} \sup_{x \in \mathcal{X}} D_s^{\varepsilon + \delta}(W(\cdot|x) \| Q) + \log \frac{1}{\delta}.$$

The first part of the proof is similar to the meta-converse of PPV [4, Thm. 31]; however, we give a conceptually simple alternative proof along the lines of Wang-Colbeck-Renner [11, Lem. 3] and Wang-Renner [12, Thm. 1].

Proof. For any code $\mathcal{C} = \{\mathcal{M}, e, d\}$ with $p_{\text{err}}(\mathcal{C}) \leq \varepsilon$ and any $Q \in \mathcal{P}(\mathcal{Y})$, the following holds.

Starting from a uniform distribution over \mathcal{M} , the Markov chain $M \xrightarrow{e} X \xrightarrow{W} Y \xrightarrow{d} M'$ induces a joint probability distribution $P_{MXYM'}$. Due to the data-processing inequality for D_h^ε , we immediately find $D_h^\varepsilon(P \times W \| P \times Q) = D_h^\varepsilon(P_{XY} \| P_X \times Q_Y) \geq D_h^\varepsilon(P_{MM'} \| P_M \times Q_{M'})$, where $P_X = P$ and $Q_{M'}$ is the distribution induced by d applied to $Q_Y = Q$.⁴ Moreover, using the test $\xi(m, m') = \delta_{m, m'}$, we can readily see that

$$\mathbb{E}_{P_{MM'}} [\xi(M, M')] = \Pr_{P_{MM'}} [M = M'] \geq 1 - \varepsilon \quad \text{and} \quad \mathbb{E}_{P_M \times Q_{M'}} [\xi(M, M')] = \frac{1}{|\mathcal{C}|}.$$

Hence, $D_h^\varepsilon(P_{MM'} \| P_M \times Q_{M'}) \geq \log |\mathcal{C}| + \log(1 - \varepsilon)$ by definition of the ε -hypothesis testing divergence. Finally, applying Lemmas 2 and 4, we find

$$\begin{aligned} \sup_{x \in \mathcal{X}} D_s^{\varepsilon + \delta}(W(\cdot|x) \| Q) &\geq D_s^{\varepsilon + \delta}(P \times W \| P \times Q) \\ &\geq D_h^\varepsilon(P \times W \| P \times Q) - \log \frac{1 - \varepsilon}{\delta} \geq \log |\mathcal{C}| - \log \frac{1}{\delta}. \end{aligned}$$

This yields the converse bound upon minimizing over $Q \in \mathcal{P}(\mathcal{Y})$. \square

⁴ Note that due to the Markov property, the encoding can be inverted probabilistically, without effecting the correlation between M and M' .

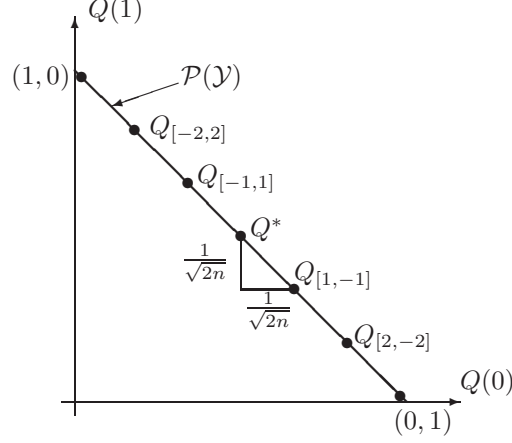


FIG. 2. Illustration of the choice of $Q_{\mathbf{k}}$ for $\mathcal{Y} = \{0, 1\}$. Note that $\zeta = 2$ for $|\mathcal{Y}| = 2$.

C. A Suitable Choice of Output Distribution Q

For n -fold repetitions of a DMC, the bound in Proposition 6 evaluates to

$$\log M^*(W^n, \varepsilon) \leq \min_{Q^{(n)} \in \mathcal{P}(\mathcal{Y}^{\times n})} \max_{\mathbf{x} \in \mathcal{X}^{\times n}} D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x}) \| Q^{(n)}) + \log \frac{1}{\delta},$$

and it thus important to find a suitable choice of $Q^{(n)} \in \mathcal{P}(\mathcal{Y}^{\times n})$ to further upper bound the above. Symmetry considerations allow us to restrict the search to distributions that are invariant under permutations of the n channel uses. Let $\zeta := |\mathcal{Y}|(|\mathcal{Y}| - 1)$ and let $\gamma > 0$ be a constant which is to be chosen later. Consider the following convex combination of product distributions:

$$Q^{(n)}(\mathbf{y}) := \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{K}} \frac{\exp(-\gamma \|\mathbf{k}\|_2^2)}{F} \prod_{i=1}^n Q_{\mathbf{k}}(y_i) + \frac{1}{2} \sum_{P_{\mathbf{x}} \in \mathcal{P}_n(\mathcal{X})} \frac{1}{|\mathcal{P}_n(\mathcal{X})|} \prod_{i=1}^n P_{\mathbf{x}} W(y_i), \quad (6)$$

where F is a normalization constant that ensures $\sum_{\mathbf{y}} Q^{(n)}(\mathbf{y}) = 1$ and

$$Q_{\mathbf{k}}(y) := Q^*(y) + \frac{k_y}{\sqrt{n\zeta}}, \quad \mathcal{K} := \left\{ \mathbf{k} \in \mathbb{Z}^{|\mathcal{Y}|} \mid \sum_y k_y = 0 \wedge k_y \geq -Q^*(y) \sqrt{n\zeta} \right\}.$$

The convex combination of $(P_{\mathbf{x}} W)^{\times n}$ in $Q^{(n)}$ is inspired partly by Hayashi [6, Thm. 2]. What we have done in our choice of $Q_{\mathbf{k}}$ is to uniformly quantize the simplex $\mathcal{P}(\mathcal{Y})$ along axis-parallel directions. The constraint that each \mathbf{k} belongs to \mathcal{K} ensures that each $Q_{\mathbf{k}}$ is a valid probability mass function. See Fig. 2. We find that

$$F \leq \sum_{\mathbf{k} \in \mathbb{Z}^{|\mathcal{Y}|}} \exp(-\gamma \|\mathbf{k}\|_2^2) = \left(\sum_{k=-\infty}^{\infty} \exp(-\gamma k^2) \right)^{|\mathcal{Y}|} \leq \left(1 + \sqrt{\frac{\pi}{\gamma}} \right)^{|\mathcal{Y}|}$$

is a finite constant. Furthermore, by construction, the representation points $\{Q_{\mathbf{k}}\}_{\mathbf{k}}$ form an ϵ -net with $\epsilon = n^{-\frac{1}{2}}$ for $\mathcal{P}(\mathcal{Y})$. Namely, for every $Q \in \mathcal{P}(\mathcal{Y})$, there exists a \mathbf{k} such that $\|Q - Q_{\mathbf{k}}\|_2 \leq n^{-\frac{1}{2}}$. This can be verified easily since by choosing a \mathbf{k} that minimizes the distance in all but one

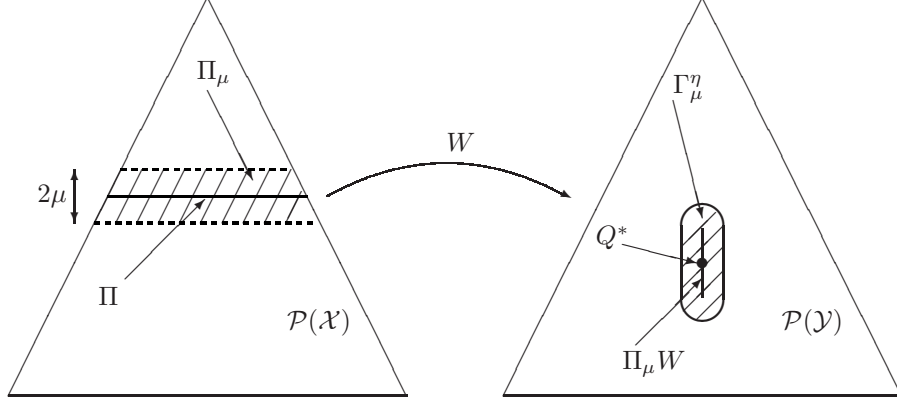


FIG. 3. Illustration of the sets in Section III D for $|\mathcal{X}| = |\mathcal{Y}| = 3$. Here, Π is not a singleton and $\Pi_\mu W$ has measure zero in $\mathcal{P}(\mathcal{Y})$ so W is rank-deficient. The unique CAOD Q^* is the image of Π under W , $\Pi_\mu W$ is the image of Π_μ under W and Γ_μ^η is the “ η -blown-up” version of $\Pi_\mu W$.

direction (say the last), yielding

$$\begin{aligned} \|Q - Q_{\mathbf{k}}\|_2^2 &= \sum_{y=1}^{|\mathcal{Y}|-1} (Q(y) - Q_{\mathbf{k}}(y))^2 + (Q(|\mathcal{Y}|) - Q_{\mathbf{k}}(|\mathcal{Y}|))^2 \\ &= \sum_{y=1}^{|\mathcal{Y}|-1} (Q(y) - Q_{\mathbf{k}}(y))^2 + \left(\sum_{y=1}^{|\mathcal{Y}|-1} Q_{\mathbf{k}}(y) - Q(y) \right)^2 \\ &\leq \sum_{y=1}^{|\mathcal{Y}|-1} \left(\frac{1}{\sqrt{n\zeta}} \right)^2 + \left(\sum_{y=1}^{|\mathcal{Y}|-1} \frac{1}{\sqrt{n\zeta}} \right)^2 = \frac{1}{n}. \end{aligned}$$

D. Continuity around the CAIDs and the unique CAOD

We will often be concerned with probability distributions close to the set of CAIDs Π in Euclidean distance, i.e., those distributions belonging to

$$\Pi_\mu := \left\{ P \in \mathcal{P}(\mathcal{X}) \mid \min_{P^* \in \Pi} \|P - P^*\|_2 \leq \mu \right\}$$

for some small $\mu > 0$. Sometimes we also need to restrict to probability distributions in Π_μ with positive conditional information variance. For a constant $v > 0$ we define

$$\Pi_\mu^v := \left\{ P \in \Pi_\mu \mid V(P, W) \geq v \right\}.$$

The image of Π_μ under W is denoted as $\Pi_\mu W$. We also consider a larger, “ η -blown-up” version, of $\Pi_\mu W$, namely

$$\Gamma_\mu^\eta := \left\{ Q \in \mathcal{P}(\mathcal{Y}) \mid \exists P \in \Pi_\mu \text{ s.t. } \|PW - Q\|_2 \leq \eta \right\}.$$

Note that $\Gamma_\mu^0 = \Pi_\mu W$ if the stochastic matrix W has full rank. See Fig. 3 for an illustration. The following Lemma summarizes known results about these sets.

Lemma 7. *Let $W : \mathcal{X} \rightarrow \mathcal{Y}$ be a DMC and $v > 0$ be a constant. There exists $\mu > 0$ and $\eta > 0$ and finite constants $V^+ > 0$, $T^+ > 0$, $q_{\min} > 0$, $\alpha > 0$, and $\beta > 0$ such that the following holds. For all $P \in \Pi_\mu$ and their projections $P^* := \arg \min_{P' \in \Pi} \|P - P'\|_2$ and all $Q \in \Gamma_\mu^\eta$ we have*

1. $Q(y) > q_{\min}$ for all $y \in \mathcal{Y}$,
2. $V(W\|Q|P) \geq \frac{V_{\min}}{2}$,
3. $I(P, W) \leq C(W) - \alpha\|P - P^*\|_2^2$,
4. $D(W\|Q|P) \leq I(P, W) + \frac{\|Q - PW\|_2^2}{q_{\min}}$,
5. $V(W\|Q|P) \leq V^+$ and $T(W\|Q|P) \leq T^+$.

Furthermore, for any $P \in \Pi_\mu^v$ we have

6. $V(W\|Q|P) \geq \frac{v}{2} > 0$,
7. $|\sqrt{V(P, W)} - \sqrt{V(P^*, W)}| \leq \beta\|P - P^*\|_2$,
8. $|\sqrt{V(W\|Q|P)} - \sqrt{V(P, W)}| \leq \beta\|Q - PW\|_2$.

Proof. Properties 1 and 2 hold for small enough μ and η by continuity since Q^* has full support [16, Cor. 1 to Thm. 4.5.1] and $V(W\|P^*W|P^*) \geq V_{\min}$. The case $V_{\min} = 0$ in Property 2 is trivial since $V(W\|Q|P) \geq 0$. Property 3 was established by Strassen [3, Eq. (4.41)] as well as PPV [4, Eq. (501)]. Since $D(W\|Q|P) = I(P, W) + D(PW\|Q)$, Property 4 follows immediately from the fact that $D(PW\|Q) \leq \frac{1}{\min_{y \in \mathcal{Y}} Q(y)}\|PW - Q\|_2^2$ (see, e.g., [17, Lem. 6.3]). Property 5 follows from the fact that $(P, Q) \mapsto V(W\|Q|P)$ and $(P, Q) \mapsto T(W\|Q|P)$ are finite and continuous on the compact set $\Pi_\mu \times \Gamma_\mu^\eta$.

Property 6 again holds for small enough η by continuity and since $V(W\|P^*W|P) \geq v$ by definition of the set Γ_μ^η . To verify Properties 7 and 8, note that the quotient $W(y|x)/Q(y) < \infty$ by Property 1. If $W(y|x)/Q(y) = 0$, the corresponding terms in the sums defining $V(P, W)$ and $V(W\|Q|P)$ are excluded because $\vartheta \log^k \vartheta \rightarrow 0$ as $\vartheta \rightarrow 0$ for all $k > 0$. Hence, $P \mapsto V(P, W)$ and $Q \mapsto V(W\|Q|P)$ are continuously differentiable on Π_μ and Γ_μ^η respectively. Because $t \mapsto \sqrt{t}$ is continuously differentiable away from 0, by Property 6, $P \mapsto \sqrt{V(P, W)}$ and $Q \mapsto \sqrt{V(W\|Q|P)}$ are Lipschitz on Π_μ and Γ_μ^η respectively. The uniformity of β in P in Property 8 can be verified by explicitly calculating the derivative of $Q \mapsto \sqrt{V(W\|Q|P)}$ and noting that it can be upper bounded by a finite constant independent of P . \square

E. Asymptotics for DMCs

We are now ready to prove our main result. Several special cases of Theorem 1 require additional proof techniques. For the convenience of the reader, we state them separately as propositions. Theorem 1 then follows as a straightforward consequence of these propositions. See Fig. 1 for a summary. The following proposition considers the “regular” case, where the channel and ε satisfy $V_\varepsilon > 0$.

Proposition 8. *For every DMC W and $\varepsilon \in (0, 1)$ such that $V_\varepsilon > 0$, the blocklength n , ε -error capacity satisfies*

$$\log M^*(W^n, \varepsilon) \leq nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + O(1).$$

Remark 4. *In the following proof of Proposition 8, we deal with all cases except $\varepsilon = \frac{1}{2}$, $V_{\min} = 0$ and $V_{\max} = V_\varepsilon > 0$. This special case will be handled in Proposition 10(i) as it uses the proof techniques in Proposition 9.*

Proof. Firstly, we employ Proposition 6 to provide a bound on $\log M^*(W^n, \varepsilon)$. We choose $\delta = n^{-\frac{1}{2}}$, which satisfies $0 < \delta < 1 - \varepsilon$ for sufficiently large n . Substitute the output distribution $Q^{(n)}$ in (6) to get

$$\log M^*(W^n, \varepsilon) \leq \max_{\mathbf{x} \in \mathcal{X}^{\times n}} \underbrace{D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x})\|Q^{(n)})}_{=: \text{cv}(\mathbf{x})} + \frac{1}{2} \log n.$$

It remains to show that each term $\text{cv}(\mathbf{x})$ in the maximization is upper bounded by $nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon) + G$ for a suitable constant G for all sufficiently large n .

We apply Lemma 7, which supplies us with finite, positive constants $\mu, \eta, V^+, T^+, q_{\min}, \alpha$ and β . If $V_{\min} > 0$, we choose $v = \frac{V_{\min}}{2}$ such that $\Pi_\mu^v = \Pi_\mu$, otherwise $v > 0$ will be specified later. See Case c) below.

We distinguish between three cases for the following; either a) \mathbf{x} satisfies $P_{\mathbf{x}} \notin \Pi_\mu$ or b) \mathbf{x} satisfies $P_{\mathbf{x}} \in \Pi_\mu^v$ or c) \mathbf{x} satisfies $P_{\mathbf{x}} \in \Pi_\mu \setminus \Pi_\mu^v$. Note that Case c) is only relevant if $V_{\min} = 0$, as otherwise $\Pi_\mu^v = \Pi_\mu$ by definition of v . This strategy in which we partition input types into such classes was proposed by Strassen [3, Sec. 4]. See also PPV [4, App. I]. Intuitively, for Case a), $P_{\mathbf{x}}$ is far from the CAIDs so the first-order term is smaller than capacity; for Case b), $P_{\mathbf{x}}$ has high conditional information variance and thus bounded skewness so we can apply the Berry-Esséen-type bound of Lemma 5 and; for Case c), $P_{\mathbf{x}}$ has small conditional information variance so we must use the Chebyshev-type bound and choose v based on V_{\max} instead of V_{\min} .

Case a): $P_{\mathbf{x}} \notin \Pi_\mu$

The mutual information outside Π_μ is bounded away from the capacity, i.e., $I(P_{\mathbf{x}}, W) \leq C' < C$ for all $P_{\mathbf{x}} \notin \Pi_\mu$.

We first apply Lemma 3 and then Lemma 5 to bound

$$\begin{aligned} \text{cv}(\mathbf{x}) &\leq D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x})\|(P_{\mathbf{x}}W)^{\times n}) + \log(2|\mathcal{P}_n(\mathcal{X})|) \\ &\leq nI(P_{\mathbf{x}}, W) + \sqrt{\frac{nV(P_{\mathbf{x}}, W)}{1 - \varepsilon - \delta}} + \log(2|\mathcal{P}_n(\mathcal{X})|). \end{aligned}$$

For the second inequality, we note that D_n in Lemma 3 evaluates to

$$D_n = \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{W(\cdot|x_i)} \left[\log \frac{W(\cdot|x_i)}{P_{\mathbf{x}}W(\cdot)} \right] = \mathbb{E}_{P_{\mathbf{x}} \times W} \left[\log \frac{W}{P_{\mathbf{x}}W} \right] = D(W\|P_{\mathbf{x}}W|P_{\mathbf{x}}) = I(P_{\mathbf{x}}, W),$$

and similar calculation can be done to show that $V_n = V(P_{\mathbf{x}}, W)$. Invoking [4, Lem. 62] and [15, Rmk. 3.1.1] yields the uniform bound $V(P_{\mathbf{x}}, W) \leq \frac{8 \log^2 e}{e^2} |\mathcal{Y}| \leq 2.3 |\mathcal{Y}|$. Hence,

$$\text{cv}(\mathbf{x}) \leq nC' + \sqrt{n} \sqrt{\frac{2.3 |\mathcal{Y}|}{1 - \varepsilon - \delta}} + (|\mathcal{X}| - 1) \log(n + 1) + \log 2.$$

Since $C' < C$, the linear term dominates the term growing with the square root of n and the term growing logarithmically in n asymptotically. Hence, it is evident that $\text{cv}(\mathbf{x}) \leq nC + \sqrt{nV_\varepsilon}\Phi^{-1}(\varepsilon)$ for sufficiently large n .

Case b): $P_{\mathbf{x}} \in \Pi_\mu^v$

For each \mathbf{x} , we denote by $Q_{\mathbf{k}(\mathbf{x})}$ the element of the ϵ -net (constructed in Section III C) closest to $P_{\mathbf{x}}W$. We note that since $\|Q_{\mathbf{k}(\mathbf{x})} - P_{\mathbf{x}}W\|_2 \leq \epsilon = n^{-\frac{1}{2}}$, we have $Q_{\mathbf{k}(\mathbf{x})} \in \Gamma_\mu^\eta$ for sufficiently large n , which enables us to apply the properties described in Lemma 7 extensively below.

We first use Lemma 3 to bound

$$\text{cv}(\mathbf{x}) \leq D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x})\|(Q_{\mathbf{k}(\mathbf{x})})^{\times n}) + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 + \log(2F).$$

We now employ Lemma 5, where we choose $P_i = W(\cdot|x_i)$ resulting in $D_n := D(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})$, $V_n := V(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})$ and $T_n := T(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})$. From Lemma 7, we have that $T_n \leq T^+$ and $0 < \frac{v}{2} < V_n \leq V^+$. We then introduce the finite constant $B := 1 + 6\sqrt{8}T_+/v^{\frac{3}{2}}$, while substituting for $\delta = n^{-\frac{1}{2}}$, to get

$$\text{cv}(\mathbf{x}) \leq nD(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}}) + \sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})} \Phi^{-1}\left(\varepsilon + \frac{B}{\sqrt{n}}\right) + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 + \log(2F).$$

We now require that $n \geq N$, where N is chosen large enough such that $\varepsilon + \frac{B}{\sqrt{N}} < 1$. This ensures that the coefficient of the term growing as \sqrt{n} in the above expression is finite. Next, we use the fact that Φ^{-1} is infinitely differentiable and $V(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}}) \leq V_+$ is finite to bound

$$\sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})} \Phi^{-1}\left(\varepsilon + \frac{B}{\sqrt{n}}\right) \leq \sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})} \Phi^{-1}(\varepsilon) + G_1.$$

for some finite constant G_1 and all $n \geq N$. Thus, defining $G_2 := G_1 + \log(2F)$, we get

$$\text{cv}(\mathbf{x}) \leq nD(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}}) + \sqrt{nV(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})} \Phi^{-1}(\varepsilon) + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 + G_2,$$

Next, we would like to replace $Q_{\mathbf{k}(\mathbf{x})}$ with $P_{\mathbf{x}}W$ in the above bound. This can be done without too much loss due to Lemma 7, which states that

$$D(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}}) \leq I(P_{\mathbf{x}}, W) + \frac{\|P_{\mathbf{x}}W - Q_{\mathbf{k}(\mathbf{x})}\|_2^2}{q_{\min}} \leq I(P_{\mathbf{x}}, W) + \frac{1}{n q_{\min}}$$

and

$$\left| \sqrt{V(W\|Q_{\mathbf{k}(\mathbf{x})}|P_{\mathbf{x}})} - \sqrt{V(P_{\mathbf{x}}, W)} \right| \leq \beta \|P_{\mathbf{x}}W - Q_{\mathbf{k}(\mathbf{x})}\|_2 \leq \frac{\beta}{\sqrt{n}}.$$

Hence, choosing $G_3 := \frac{1}{q_{\min}} + \beta|\Phi^{-1}(\varepsilon)| + G_2$, we find that

$$\text{cv}(\mathbf{x}) \leq nI(P_{\mathbf{x}}, W) + \sqrt{nV(P_{\mathbf{x}}, W)} \Phi^{-1}(\varepsilon) + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 + G_3.$$

In the following, we use the fact that all distributions (and types) $P_{\mathbf{x}}$ in Π_{μ} satisfy $I(P_{\mathbf{x}}, W) \leq C - \alpha\xi^2$ and $|\sqrt{V(P_{\mathbf{x}}, W)} - \sqrt{V(P^*, W)}| \leq \beta\xi$, where $P^* := \arg \min_{P' \in \Pi} \|P_{\mathbf{x}} - P'\|_2$ (which is unique) and $\xi := \|P_{\mathbf{x}} - P^*\|_2$. Hence,

$$\text{cv}(\mathbf{x}) \leq nC + \sqrt{nV(P^*, W)} \Phi^{-1}(\varepsilon) + \left(-\alpha\xi^2 n + \beta|\Phi^{-1}(\varepsilon)|\xi\sqrt{n} + \gamma\|\mathbf{k}(\mathbf{x})\|_2^2 \right) + G_3. \quad (7)$$

It thus remains to show that the term in the bracket is upper bounded by a constant, for an appropriate choice of γ . Let $\|W\|_2 := \max\{\|\mathbf{u}W\|_2 \mid \|\mathbf{u}\|_2 \leq 1\}$ be the spectral norm of the matrix W . It is easy to see that $\|W\|_2 \leq \sqrt{|\mathcal{X}|}$. From the construction of the ϵ -net in Section III C,

$$\begin{aligned} \|\mathbf{k}(\mathbf{x})\|_2 &= \sqrt{n\zeta} \|Q_{\mathbf{k}(\mathbf{x})} - Q^*\|_2 \\ &\leq \sqrt{n\zeta} \left(\|Q_{\mathbf{k}(\mathbf{x})} - P_{\mathbf{x}}W\|_2 + \|P_{\mathbf{x}}W - Q^*\|_2 \right) \\ &\leq \sqrt{n\zeta} \left(\frac{1}{\sqrt{n}} + \|W\|_2 \xi \right). \end{aligned}$$

Substituting this bound into (7), we find that the term in the bracket evaluates to

$$(\gamma\zeta\|W\|_2^2 - \alpha)\xi^2n + (\beta|\Phi^{-1}(\varepsilon)| + 2\gamma\zeta\|W\|_2)\xi\sqrt{n} + \gamma\zeta$$

The expression is a quadratic polynomial in $\xi\sqrt{n}$ and has a finite maximum if we choose γ such that $\gamma\zeta\|W\|_2^2 < \alpha$. Hence, we can write

$$\text{cv}(\mathbf{x}) \leq nC + \sqrt{nV(P^*, W)}\Phi^{-1}(\varepsilon) + G_4$$

for an appropriate constant G_4 and $n \geq N$.

$$\text{Case c) } P_{\mathbf{x}} \in \Pi_{\mu} \setminus \Pi_{\mu}^v$$

Note that this case only appears if $V_{\min} = 0$, $V_{\max} = V_{\varepsilon} > 0$ and $\varepsilon \geq \frac{1}{2}$. We consider the case $\varepsilon > \frac{1}{2}$ (cf. Remark 4) leaving the $\varepsilon = \frac{1}{2}$ case for Proposition 10(i). We have

$$\begin{aligned} \text{cv}(\mathbf{x}) &\leq D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x})\|(P_{\mathbf{x}}W)^{\times n}) + \log(2|\mathcal{P}_n(\mathcal{X})|) \\ &\leq nI(P_{\mathbf{x}}, W) + \sqrt{\frac{nV(P_{\mathbf{x}}, W)}{1-\varepsilon-\delta}} + \log(2|\mathcal{P}_n(\mathcal{X})|) \\ &\leq nI(P_{\mathbf{x}}, W) + \sqrt{\frac{nv}{1-\varepsilon-\delta}} + \log(2|\mathcal{P}_n(\mathcal{X})|) \end{aligned}$$

Now we choose $v > 0$ to be any constant satisfying

$$\sqrt{\frac{v}{1-\varepsilon-\delta}} + \frac{\log(2|\mathcal{P}_n(\mathcal{X})|)}{\sqrt{n}} \leq \sqrt{V_{\max}}\Phi^{-1}(\varepsilon).$$

It is certainly possible to find such a v since the number of types is polynomial so δ and the second term on the left are arbitrarily small for large enough n . Furthermore, $\sqrt{V_{\max}}\Phi^{-1}(\varepsilon) > 0$. This is where $\varepsilon \neq \frac{1}{2}$ is crucial. Uniting the preceding two bounds yields

$$\text{cv}(\mathbf{x}) \leq nI(P_{\mathbf{x}}, W) + \sqrt{nV_{\max}}\Phi^{-1}(\varepsilon) \leq nC + \sqrt{nV_{\max}}\Phi^{-1}(\varepsilon).$$

Summarizing the bounds for Cases a), b) and c), we thus have the following asymptotic expansion for all n sufficiently large:

$$\begin{aligned} \log M^*(W^n, \varepsilon) &\leq \max_{P^* \in \Pi} nC + \sqrt{nV(P^*, W)}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + G_4 \\ &= nC + \sqrt{nV_{\varepsilon}}\Phi^{-1}(\varepsilon) + \frac{1}{2}\log n + G_4, \end{aligned}$$

where the last equality follows by definition of V_{ε} . \square

Surprisingly, the first order approximation is accurate up to a constant term if $V_{\varepsilon} = 0$ unless the channel is exotic and $\varepsilon \geq \frac{1}{2}$.

Proposition 9. *For every DMC W and $\varepsilon \in (0, 1)$ such that $V_{\varepsilon} = 0$, the blocklength n , ε -error capacity satisfies $\log M^*(W^n, \varepsilon) \leq nC + O(1)$, unless the channel is exotic and $\varepsilon \geq \frac{1}{2}$.*

Proof. Again, from our bound on the converse for general channels (Prop. 6), we have

$$\log M^*(W^n, \varepsilon) \leq \max_{\mathbf{x} \in \mathcal{X}^{\times n}} \underbrace{D_s^{\varepsilon+\delta}(W^n(\cdot|\mathbf{x})\|Q^{(n)})}_{=:\text{cv}(\mathbf{x})} + \log \frac{1}{\delta}. \quad (8)$$

For this analysis, we ignore Section III C and instead choose $Q^{(n)} = (Q^*)^{\times n}$ to be the n -fold extension of the CAOD. We choose $\delta = \frac{1}{2} - \varepsilon$ if $\varepsilon < \frac{1}{2}$ and $\delta = \frac{1-\varepsilon}{2}$ otherwise; hence, the term $\log \frac{1}{\delta}$ is finite and independent of n . Since F is bounded and $\mathbf{k} = \mathbf{0}$ for $(Q^*)^{\times n}$, it remains to prove that $\text{cv}(\mathbf{x}) \leq nC + O(1)$ for all \mathbf{x} .

For this purpose, let $m(\mathbf{x})$ be the number of non-zero variance letters in \mathbf{x} , i.e., $m(\mathbf{x}) := nP_{\mathbf{x}}(\mathcal{X}_+) = \sum_{i=1}^n 1_{\{x_i \in \mathcal{X}_+\}}$ where $\mathcal{X}_+ := \{x \in \mathcal{X} : V(W(\cdot|x)\|Q^*) > 0\}$. There exist finite constants v_{\min}, v_{\max} and t_{\max} such that, for every $x \in \mathcal{X}_+$,

$$0 < v_{\min} \leq V(W(\cdot|x)\|Q^*) \leq v_{\max}, \quad \text{and} \quad T(W(\cdot|x)\|Q^*) \leq t_{\max}.$$

By the definition of $D_n := D(W\|Q^*|P_{\mathbf{x}})$, $V_n := V(W\|Q^*|P_{\mathbf{x}})$ and $T_n := T(W\|Q^*|P_{\mathbf{x}})$ (cf. Lemma 5), we have

$$\frac{m(\mathbf{x})}{n} v_{\min} \leq V_n \leq \frac{m(\mathbf{x})}{n} v_{\max}, \quad \text{and} \quad T_n \leq \frac{m(\mathbf{x})}{n} t_{\max}. \quad (9)$$

Further defining $B_n := 6 T_n / V_n^{\frac{3}{2}}$, we thus find

$$B_n \leq \sqrt{\frac{n}{m(\mathbf{x})}} L \quad \text{where} \quad L := \frac{6 t_{\max}}{v_{\min}^{3/2}} < \infty.$$

Let m^* be an integer satisfying $L/\sqrt{m^*} \leq r'$ where r' is chosen such that $\Phi^{-1}(\frac{1}{2} + r) \leq 3r$ for all $r \in [0, r']$. The choice $r' = 0.35$ does the job.

For $\varepsilon < \frac{1}{2}$, following Strassen's argument [3, Eq. (4.53)-(4.54)] (see also PPV [4, App. I]), we distinguish between two classes of sequences as follows: the sequence \mathbf{x} satisfies either a) $m(\mathbf{x}) \geq m^*$, or b) $m(\mathbf{x}) < m^*$. Finally, c) considers the case where W is not exotic and $\varepsilon \geq \frac{1}{2}$. Intuitively, for Case a), we can use the Berry-Esséen-type bound because $m(\mathbf{x})$ is large, and hence B_n can be bounded appropriately; for Case b), we use the Chebyshev-type bound because $m(\mathbf{x})$ is small and; for Case c), we use the non-exoticness of W to bound D_n far away from C .

Case a): $\varepsilon < \frac{1}{2}$ and $m(\mathbf{x}) \geq m^$*

We apply the Berry-Esséen-type bound in Lemma 5 to (8) to find

$$\begin{aligned} \text{cv}(\mathbf{x}) &\leq nD_n + \sqrt{nV_n} \Phi^{-1} \left(\varepsilon + \delta + \frac{B_n}{\sqrt{n}} \right) \\ &\leq nD_n + \sqrt{nV_n} \Phi^{-1} \left(\frac{1}{2} + \frac{L}{\sqrt{m(\mathbf{x})}} \right) \leq nD_n + 3L \sqrt{\frac{nV_n}{m(\mathbf{x})}}. \end{aligned} \quad (10)$$

Here, we used the fact that $\varepsilon + \delta = \frac{1}{2}$ by definition of δ and the proof concludes with the observation that $\frac{nV_n}{m(\mathbf{x})} \leq v_{\max}$ is bounded by a constant, and $D_n \leq C$ for all \mathbf{x} .

Case b): $\varepsilon < \frac{1}{2}$ and $m(\mathbf{x}) < m^$*

We use the Chebyshev-type bound in Lemma 5 to (8) yielding

$$\text{cv}(\mathbf{x}) \leq nD_n + \sqrt{\frac{nV_n}{1-\varepsilon-\delta}} = nD_n + \sqrt{2nV_n}. \quad (11)$$

Since by (9), $nV_n \leq m^* v_{\max}$ and $D_n \leq C$ for all \mathbf{x} , we find the desired bound.

Case c): not exotic, $\varepsilon \geq \frac{1}{2}$

Lemma 5 applied to (8) again yields

$$\text{cv}(\mathbf{x}) \leq nD_n + \sqrt{\frac{nV_n}{1-\varepsilon-\delta}} = nD_n + \sqrt{\frac{2nV_n}{1-\varepsilon}},$$

because in this case, $\delta = \frac{1-\varepsilon}{2}$. By virtue of the fact that $V_{\max} = 0$ and W is not exotic, we have that either

$$D(W(\cdot|x)\|Q^*) < C \quad \text{or} \quad V(W(\cdot|x)\|Q^*) = 0 \quad (12)$$

for all symbols $x \in \mathcal{X}$. If \mathcal{X}_+ is empty, we have $V_n = 0$ and the bound is immediate. Otherwise, we define $\psi := C - \max_{x \in \mathcal{X}_+} D(W(\cdot|x)\|Q^*) > 0$, which is positive due to the condition in (12).

Using this, we find that $nD_n \leq nC - m(\mathbf{x})\psi$ and $nV_n \leq v_{\max}m(\mathbf{x})$ by (9). Thus,

$$\text{cv}(\mathbf{x}) \leq nC - m(\mathbf{x})\psi + \sqrt{\frac{2m(\mathbf{x})v_{\max}}{1-\varepsilon}}$$

The latter two terms constitute a quadratic polynomial in $\sqrt{m(\mathbf{x})}$, and hence, their sum has a finite maximum. \square

Finally, we deal with the case that was left out in Proposition 8.

Proposition 10. *Let $\varepsilon = \frac{1}{2}$. The following hold:*

- (i) *For every DMC W such that $V_{\min} = 0$ and $V_{\max} > 0$, the blocklength n , ε -error capacity satisfies $\log M^*(W^n, \varepsilon) \leq nC + \frac{1}{2} \log n + O(1)$.*
- (ii) *For every exotic DMC W (in particular, $V_{\max} = 0$), the same bound as in (i) holds.*

Proof. By placing no assumptions on $V_{\max} \geq 0$, we can prove both parts in tandem. The proof follows closely that of Proposition 9 with the exception that we choose $\delta = n^{-\frac{1}{2}}$ so the $\log \frac{1}{\delta}$ term evaluates to $\frac{1}{2} \log n$. It remains to show that $\text{cv}(\mathbf{x}) \leq nC + O(1)$. We split the analysis into Cases a) and b) as in Proposition 9 and let $D_n := D(W\|Q^*|P_{\mathbf{x}})$ and $V_n := V(W\|Q^*|P_{\mathbf{x}})$.

Case a): $\varepsilon = \frac{1}{2}$, $V_{\min} = 0$ and $m(\mathbf{x}) \geq m^*$

By the same steps that led to (10), we have

$$\text{cv}(\mathbf{x}) \leq nD_n + 3(L+1)\sqrt{\frac{nV_n}{m(\mathbf{x})}}$$

because $\delta = n^{-\frac{1}{2}}$. We obtain the desired bound by noting that $\frac{nV_n}{m(\mathbf{x})} \leq v_{\max}$ and $D_n \leq C$.

Case b): $\varepsilon = \frac{1}{2}$, $V_{\min} = 0$ and $m(\mathbf{x}) < m^*$

By the same steps that led to (11), we have

$$\text{cv}(\mathbf{x}) \leq nD_n + \sqrt{4nV_n}$$

because $1-\varepsilon-\delta = \frac{1}{2}-\delta \geq \frac{1}{4}$ for all $n \geq 4$. The proof is completed by noting that $nV_n \leq m^*v_{\max}$ and $D_n \leq C$. \square

Proof of Theorem 1. The first statement follows by Propositions 8 and 10(i). The second statement follows by Proposition 9. \square

Acknowledgements

MT thanks Ligong Wang for helpful explanations. VYFT thanks Yanina Shkel for insightful discussions and Pierre Moulin for sharing his ITA paper [10]. MT is supported by the National Research Foundation and the Ministry of Education of Singapore. VYFT would like to acknowledge funding support from A*STAR, Singapore.

-
- [1] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. Journal*, 27:379–423, 1948.
 - [2] J. Wolfowitz. *Coding Theorems of Information Theory*. Springer-Verlag, New York, 3rd edition, 1978.
 - [3] V. Strassen. Asymptotische Abschätzungen in Shannons Informationstheorie. In *Trans. Third Prague Conf. Inf. Theory*, pages 689–723, Prague, 1962.
 - [4] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding in the finite blocklength regime. *IEEE Trans. on Inf. Th.*, 56:2307–59, May 2010.
 - [5] Y. Polyanskiy. *Channel coding: Non-asymptotic fundamental limits*. PhD thesis, Princeton University, 2010.
 - [6] M. Hayashi. Information spectrum approach to second-order coding rate in channel coding. *IEEE Trans. on Inf. Th.*, 55:4947–66, Nov 2009.
 - [7] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.
 - [8] P. Moulin. The log-volume of optimal constant-composition codes for memoryless channels, within $O(1)$ bits. In *Int. Symp. Inf. Th.*, Cambridge, MA, 2012.
 - [9] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley and Sons, 2nd edition, 1971.
 - [10] P. Moulin. The log-volume of optimal codes for memoryless channels, up to a few nats. In *Info. Th. and Appl. Workshop*, San Diego, CA, 2012.
 - [11] L. Wang, R. Colbeck, and R. Renner. Simple channel coding bounds. In *Intl. Symp. Inf. Th.*, Seoul, South Korea, 2009.
 - [12] L. Wang and R. Renner. One-shot classical-quantum capacity and hypothesis testing. *Physical Review Letters*, 108:200501, May 2012.
 - [13] M. Tomamichel and M. Hayashi. A hierarchy of information quantities for finite block length analysis of quantum tasks. [arXiv:1208.1478 \[quant-ph\]](#), Sep 2012.
 - [14] F. Dupuis, L. Krämer, P. Faist, J. M. Renes, and R. Renner. Generalized entropies. [arXiv:1211.3141 \[quant-ph\]](#), Nov 2012.
 - [15] T. S. Han. *Information-Spectrum Methods in Information Theory*. Springer Berlin Heidelberg, Feb 2003.
 - [16] R. G. Gallager. *Information Theory and Reliable Communication*. Wiley, New York, 1968.
 - [17] I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. on Inf. Th.*, 52(3):1007–16, Mar 2006.